COPY RIGHT

ELSEVIER
SSRN

Title M²-CTTS: END-TO-END MULTI-SCALE MULTI-MODAL CONVERSATIONAL TEXT-TO-SPEECH SYNTHESIS

Paper Authors **Teresa Matoso Manguangua Victor, Miguel João Manuel Filho, Alan Claude Award, Kunal, Rishi Matura, Hitesh**

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

M²-CTTS: END-TO-END MULTI-SCALE MULTI-MODAL CONVERSATIONAL TEXT-TO-SPEECH SYNTHESIS

Teresa Matoso Manguangua Victor[1], Miguel João Manuel Filho[1], Alan Claude Award[2], Kunal[3], Rishi Matura[3], Hitesh[3]

[1]Department of Engineering and Technology, Instituto Superior Politécnico de Tecnologias e Ciências (ISPTEC), Republic of Angola

[2]School of Biologia, Newcastle University, United Kingdom

[3]Department of Computer science Engineering, Chandigarh University, India

Corresponding Author: Teresa Matoso Manguangua Victor (teresa.victor@isptec.co.ao)

ABSTRACT

Conversational text-to-speech (TTS) aims to synthesize speech with proper prosody of reply based on the historical conversation. However, it is still a challenge to comprehensively model the conversation, and a majority of conversational TTS systems only focus on extracting global information and omit local prosody features, which contain important fine-grained information like keywords and emphasis. More- over, it is insufficient to only consider the textual features, and acoustic features also contain various prosody information. Hence, we propose M²-CTTS, an end- to-end multi-scale multi-modal conversational text-to- speech system, aiming to comprehensively utilize historical conversation and enhance prosodic expression. More specifically, we design a textual context module and an acoustic context module with both coarse-grained and fine- grained modeling. Experimental re- sults demonstrate that our model mixed with fine-grained context information and additionally considering acoustic features achieves better prosody performance and naturalnessin CMOS tests.

Index Terms— speech synthesis, conversational TTS, prosody, multi-grained, multi-modal

1. INTRODUCTION

In the past few years, deep learning based text-to-speech (TTS) models have developed rapidly and they [1, 2] can produce natural speech with narrator-like high quality that matches human levels. Moreover, by training on large-scale emotional speech corpus, many expressive speech synthesis systems have been proposed and they [3, 4, 5, 6, 7] can produce speech with rich prosody and expressiveness. Conversational TTS, as a more complex technique in the field of Human-Computer Interaction (HCI), aims to synthesize speech with suitable prosody according to history conversation. Those expressive speech synthesis methods cannot perform well if they are directly applied to conversational

TTS [8], because conversational TTS relies heavily on historical conversation and it is crucial to model such dialogue information.Some conversation modeling methods have been pro posed recently. Guo et al. [9] proposed a context encoder to reinforce the utterance-level information of history context from textual aspect, and an extra auxiliary encoder is used to extract useful statistical text features like semantic and syntactic features. Cong et al. [10] introduced an extra acoustic context encoder to acquire the acoustic embedding from previous speech. Besides, Li et al. [8] proposed a graph-based multi-modal context modeling method to model inter-speaker and intra-speaker influence in conversations.However, these researches in

conversational TTS only fo- cus on predicting prosody embedding from the global level within single modality and neglect to consider local features from different aspects, which are insufficient to reflect actual human behaviors [11] like keywords emphasis or changes of prosody. Hence, we adopt a hierarchical modeling method, including utterance-level and phoneme-level modules from both textual and acoustic aspects. Moreover, some meth- ods [9, 10] simply use concatenation operation and GRU for modeling, which has a limited ability to describe relation- ship between current and past utterances. Hence, we enhance the connection by multi-head cross-modal attention mecha- nism in our work. Besides, we further apply next predictor module for better prosody similarity and use Style-Adaptive Layer Normalization (SALN) [5] to infuse prosody informa- tion. We conduct experiments on conversational speech cor- pus DailyTalk [12] and the results show that our method out- performs other methods in terms of CMOS test. Especially, fine-grained expressiveness has been greatly strengthened.

Our major contributions are listed in the following. (1) we not only extract global features but also make use of local information to enhance expressiveness, and (2) we develop a method that uses both textual and acoustic conversation in- formation, and (3) we better utilize prosody embeddings by applying SALN and prosody predictor module.

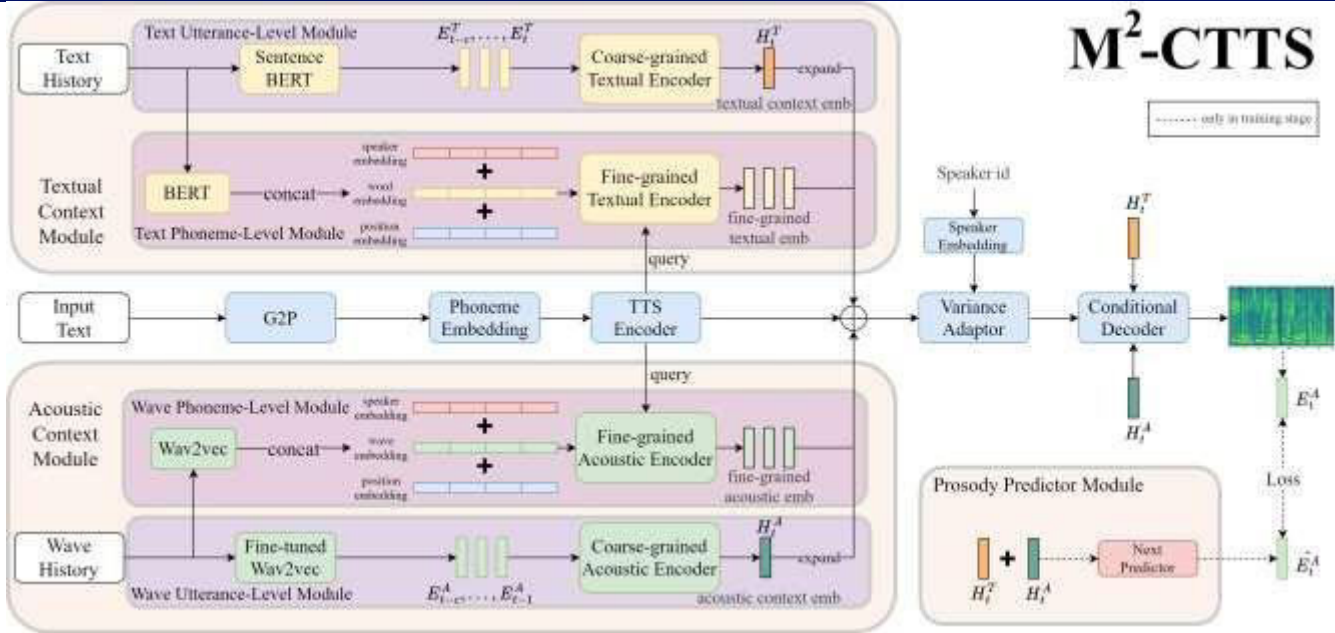The rest of the paper is organized as follows. Section 2

Fig. 1. Overview of our proposed M^2-CTTS model.

introduces our proposed conversational TTS system in detail. Section 3 shows our experiment results and discussions. The paper is concluded in Section 4.

2. M^2-CTTS MODEL

The entire model architecture of M^2-CTTS is depicted in Fig. 1. M^2-CTTS is based on FastSpeech2 [2] which consists of an encoder, a decoder and a variance adaptor. In order to capture both textual and acoustic information from the history conversation, we design a Textual Context Module (TCM) and an Acoustic Context Module (ACM). Furthermore, in order to consider hierarchical history context at the same time, we design a Text Utterance-Level Module (TUM) and a Text Phoneme-Level Module (TPM) in TCM and design a Wave Utterance-Level Module (WUM) and a Wave Phoneme-Level Module (WPM) in ACM. Moreover, we use Conditional Decoder and Prosody Predictor Module (PPM) to enhance the expressiveness of the generated speech. Specifically, we adopt Style-Adaptive Layer Normalization (SALN) [5] in decoder as Conditional Decoder to increase the expressiveness and better utilize the history context.

2.1. Multi-scale Dialogue Modeling

Although global information like emotions or intentions allows the model to understand the conversation, detailed information like keywords or emphasis that include strong or implicit meanings should also be highlighted. Besides, inconsistent or mixed style within utterances may lead to unnatural prosody. We also empirically found that fixed-length conversation embedding worked poorly and led to average prosody. Hence, we additionally combine coarse-grained with fine-grained conversation modeling and adopt multihead attention to acquire comprehensive history information. We denote the conversation history at t time stamp as:

$$Conversation = \{A_{t-c}, B_{t-c+1}, \ldots, A_{t-1}, B_t\} \qquad (1)$$

where c is a memory capacity parameter that determines how many turns are taken into consideration between two speakers A and B.

Coarse-grained Context Modeling. In terms of coarse-grained context modeling, we utilize Text Utterance-Level Module (TUM) and Wave Utterance-Level Module (WUM) to acquire global context information.

For TUM, we follow and modify the module proposed in [9] to model the textual utterance-level information, which consists of a Sentence BERT [13] based textual utterance-level feature extractor and a Coarse-grained Textual Encoder. Because Sentence BERT can capture semantic information, we adopt it to extract history sentences embeddings $E^T_{t-c:t-1}$ from $Conversation$ with last c turns dialogue and current utterance embedding E^T_t For Coarse-grained Textual Encoder, a GRU layer is used to encode the history embeddings $E^T_{t-c:t-1}$ first for considering cumulative past conversational information. Then we concatenate the final hidden states of GRU with the current utterance embedding E^T_t and feed them into a linear layer with additional attention mechanism to calculate the weighted global textual context embedding H^T_t as shown in Fig. 1.

For WUM, we adopt a similar architecture to extract the

global acoustic context embedding H_t^A, except for the feature extractor and we do not use the speech of the current utterance B_t. Instead of the conventional methods that use spectral-based features and a reference encoder or Global Style Token (GST) [3], we adopt a Wav2vec 2.0 [14] fine-tuned in the downstream Speech Emotion Recognition (SER) task to extract the acoustic utterance-level prosody information E_i^A directly from raw audio. This method brings twofold benefits. The pretrained model can capture more comprehensive acoustic features in contrast to spectral-based features which lack phase information and contain less prosodic data than raw audio. In addition, using pretrained model can mitigate the training difficulty of complex models and have robust performance.

Fine-grained Context Modeling Although global information like emotions or intentions allows the model to understand the conversation, detailed information like keywords or emphasis that include strong or implicit meanings should also be highlighted.

Considering a real-world conversation, people will react to specific words or phrases that others say in communication. Therefore, in order to simulate these situations, we develop a Text Phoneme-level Module (TPM) and a Wave Phoneme-Level Module (WPM) to model fine-grained history information.

In TPM and WPM, hidden feature sequence P_i from i-th sentence or utterance is acquired from BERT [15] and Wav2vec 2.0 [14] respectively. Multi-head cross-modal attention mechanism is then utilized to identify essential phrases or frames which have an important impact on context understanding and expressiveness. A more detailed introduction is below.

2.2. Multi-modal Context Modeling

In addition to semantic and syntactic features, paralinguistic information like acoustic prosodic cues also play an important role in context understanding, since people speak in different tones even in the same contents or situations. Thus, modeling both modalities is significant.

Textual Context Modeling. As shown in Fig. 1, Textual Context Module contains TPM and TUM. TPM consists of a feature extractor and a Fine-grained Textual Encoder. We adopt BERT [15] to acquire the hidden feature sequences P_i from i-th sentence and aggregate all last c dialogues into a long sequence. After that, we add the speaker ID to reserve speaker identity and we also use sinusoidal positions to represent the conversation orders. In terms of Fine-grained Textual Encoder, 1D convolutional layer is first adopted for contextualization. The output of TTS encoder is then used as the query for a multi-head cross-attention module with historical fine-grained representation sequences as both key and value. The resulting weighted context representation will be added back to the output of TTS encoder with the residual design. The structure of TUM is depicted in Sec. 2.1. Thus, the important

keywords or phrases from the textual aspect are considered.

Acoustic Context Modeling. As shown in Fig. 1, Acoustic Context Module includes WPM and WUM. WPM consists of a feature extractor and a Fine-grained Acoustic Encoder. Different from using mel spectrogram as acoustic input, we adopt Wave2vec 2.0 [14] to extract the hidden feature sequences from raw audios. The other processes are similar to TPM described above. Therefore, the local expression or emphasis from the acoustic perspective is noticed.

2.3. Constraint of Prosody

Similar to [10], we additionally utilize a Prosody Predictor Module (PPM) to predict prosody embedding of current utterance E_t^A from multi-modal global context embeddings H_t^T and H_t^A. As shown in Eq. 2, an additional MSE loss is adopted to constrain coarse-grained context modules to predict the prosody expression of current utterance based on the history.

$$\hat{E}_t^A = PPM(H_t^T, H_t^A), \; Loss = ||\hat{E}_t^A, E_t^A|| \quad (2)$$

where $\hat{E}_t^A$ stands for predicted prosody embedding. E_t^A is the ground-true prosody embedding extracted by fine-tuned Wav2vec model. It should be noted that the prosody predictor module and loss computation are only used during training.

3. EXPERIMENTS

3.1. Experimental Setup

We use public English corpus DailyTalk [16] to conduct our experiments. The DailyTalk dataset is derived from DailyDialog dataset [16] and it contains 2,541 dialogues performed by one male and one female speaker. Every dialogue in DailyTalk has more than five turns. For convenience, they are split into 23,773 audio clips by dialogue turns and they contain 20 hours in total. We leave 128 dialogues for validation set and others for training set. In our experiments, all utterances are down-sampled to 22050Hz and are used to extract 80 dimensional mel spectrogram.

We use FastSpeech2 as our TTS backbone and we adopt unsupervised duration modeling [17] rather than using supervised duration model with external aligner, such as Montreal Forced Aligner [18], because external aligners have risk of out-of-distribution problem and soft alignment has more flexibility for expressive performance. We use the pretrained HiFi-GAN [19] model as our vocoder to convert the 80 dimensional mel spectro-grams to 22050Hz audio files. Our conversational TTS models are all trained for 400K steps with a batch size of 16 on GeForceRTX 3090. The training hyper-parameters such as optimizers are based on the original configuration.

For textual feature extractors, we use pretrained Sentence BERT [1] in TUM, and use the original BERT in TPM. For

Table 1. The comparison methods and the MOS results.

Methods	TUM	WUM	TPM	WPM	MOS Score
M0	reconstructed from ground-truth				4.38 ± 0.06
M1					3.64 ± 0.07
M2	✓				3.71 ± 0.06
M3		✓			3.65 ± 0.07
M4	✓		✓		3.64 ± 0.07
M5		✓		✓	3.70 ± 0.07
M6	✓	✓			3.73 ± 0.06
M7	✓	✓	✓	✓	3.74 ± 0.07

acoustic feature extractors, we adopt pretrained Wav2vec 2.0 [2] fine-tuned on IEMOCAP [20] emotional training data in WUM, and use original Wav2vec 2.0 in WPM.

To study the performance of our proposed method, we use the comparison methods M0 to M7 for evaluation. Detailed combinations are in Table 1. M1 is the FastSpeech2 baseline model and M2 to M7 all adopt Conditional Decoder and prosody Predictor Module to enhance the expressiveness.

3.2. Evaluation

We conduct comparison mean opinion score (CMOS) test to evaluate the performance of each designed pair (A, B) in test set. We randomly select 20 synthesized utterances at i-th turns U_i and four ground-truth history contexts with both text and speech $U_{i-4:i-1}$ for reference to compare the naturalness and expressiveness. Especially, participants need to judge whether the prosody is more expressive and more suitable with the history dialogue, and rate on a scale from -3 (Completely A model better) to 3 (Completely B model better) with 1-point discrete increments according to the criteria. The absolute value shows the degree.

The CMOS results are shown in Table 2. We find that using TUM or WUM to model dialogue history in utterance-level from textual or acoustic aspects respectively can both improve the expressiveness a lot. We also find that combining fine-grained features in acoustic aspect has significant benefits in synthesizing speech with rich prosody. However, we do not find similar conclusion in textual aspect. We suppose that local acoustic features contain more implicit information, since people utter the same word or phrase with a variety of prosody and emotion in different situations. Comparison "M4 vs. M5" considering multi-scale information but from different aspects also confirms that acoustic features is more effective. Furthermore, we investigate the performance of adopting both textual and acoustic modalities. CMOS score of "M2 vs. M6" indicates that combining two aspects in utterance level can just slightly enhance the performance. We then visualize the attention and find that both TUM and WUM focus on the same utterances in most cases, especially the latest turn. As

Table 2. The results of the CMOS tests. CMOS describes the preference degree between A and B (denote as A vs. B). The Preference is calculated according to the CMOS score.

		Preference(%)		
	CMOS	Left	Neutral	Right
M1 vs. M2	0.25	18.9	46.1	35.0
M1 vs. M3	0.19	14.4	53.9	31.7
M2 vs. M4	0.03	25.6	45.5	28.9
M3 vs. M5	0.21	17.2	46.1	36.7
M4 vs. M5	0.19	18.9	45.5	35.6
M2 vs. M6	0.05	25.6	43.8	30.6
M2 vs. M7	0.34	17.8	38.3	43.9
M6 vs. M7	0.32	16.7	41.1	42.2

a consequence, the extracted information has high degree of similarity. The test cases "M2 vs. M7" and "M6 vs. M7" show that our method is superior to other methods [9, 10] respectively proposed in earlier research. M2-CTTS has better understanding of the history conversation and greatly enhances the fine-grained prosodic expression.

We also conduct a five-scale mean opinion score (MOS) test to evaluate the quality and naturalness of the speech. The greater the score, the better the quality. Results in Table 1 show that modeling the prosody information from different aspects can relieve the problem of generating average prosody. We also notice that despite our method containing many modules, it does not necessarily impair the quality and naturalness thanks to the prior knowledge from pretrained models. Examples of synthesized speech can be found on the project page [3].

4. CONCLUSION

In this paper, we propose an end-to-end multi-scale multi-modal conversational speech synthesis system (M^2-CTTS) that models multi-grained context information extracted from both acoustic and textual features. Experimental results show that considering both acoustic and textual modalities can enhance the prosody and naturalness of synthesized speech. Besides, combining coarse-grained features with fine-grained features will further improve expressiveness. especially the fine-grained prosody extracted from acoustic features. For future work, we will investigate the fine-grained modeling of conversational TTS with controllability.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *international conference on machine learning*. PMLR, 2018, pp. 4693–4702.

[2] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[3] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[4] Younggun Lee and Taesu Kim, "Robust and fine-grained prosody control of end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5911–5915.

[5] Dongchan Min, Dong Bok Lee, Eunho Yang, and Sung Ju Hwang, "Meta-stylespeech: Multi-speaker adaptive text-to-speech generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 7748–7759.

[6] Pengfei Wu, Junjie Pan, Chenchang Xu, Junhui Zhang, Lin Wu, Xiang Yin, and Zejun Ma, "Cross-speaker emotion transfer based on speaker condition layer normalization and semi-supervised training in text-to-speech," *arXiv preprint arXiv:2110.04153*, 2021.

[7] Yi Lei, Shan Yang, Xinsheng Wang, and Lei Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 853–864, 2022.

[8] Jingbei Li, Yi Meng, Chenyi Li, Zhiyong Wu, Helen Meng, Chao Weng, and Dan Su, "Enhancing speaking styles in conversational text-to-speech synthesis with graph-based multi-modal context modeling," in *ICASSP*. 2022, pp. 7917–7921, IEEE.

[9] Haohan Guo, Shaofei Zhang, Frank K. Soong, Lei He, and Lei Xie, "Conversational end-to-end TTS for voice agents," in *IEEE Spoken Language Technology Workshop, SLT 2021, Shenzhen, China, January 19-22, 2021*. 2021, pp. 403–409, IEEE.

[10] Jian Cong, Shan Yang, Na Hu, Guangzhi Li, Lei Xie, and Dan Su, "Controllable context-aware conversational speech synthesis," in *Interspeech*. 2021, pp. 4658–4662, ISCA.

[11] Xin Xie, Andrés Buxó-Lugo, and Chigusa Kurumada, "Encoding and decoding of meaning through structured variability in intonational speech prosody," *Cognition*, vol. 211, pp. 104619, 2021.

[12] Keon Lee, Kyumin Park, and Daeyoung Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," *arXiv preprint arXiv:2207.01063*, 2022.

[13] Nils Reimers and Iryna Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[14] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework forself-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[16] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

[17] Rohan Badlani, Adrian Lańcucki, Kevin J Shih, Rafael Valle, Wei Ping, and Bryan Catanzaro, "One tts alignment to rule them all," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6092–6096.

[18] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi.," in *Interspeech*, 2017, vol. 2017, pp. 498–502.

[19] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17022–17033, 2020.

[20] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.